

# Existence of SBP operators with diagonal norm

Nathan Albin<sup>a,\*</sup>, Joshua Klarmann<sup>a</sup>

<sup>a</sup>*Department of Mathematics, Kansas State University, 138 Cardwell Hall, Manhattan, KS 66506*

---

## Abstract

Although Summation by Parts (SBP) operators with diagonal energy norm and interior order of accuracy as high as 10 are present in the literature, there is currently no general existence theory. Moreover, the standard technique for constructing such operators involves the solution of a nonlinear system of equations by Computer Algebra System (CAS), which may fail to locate some solutions. This paper presents a new algorithm for determining the existence or nonexistence of diagonal-norm first derivative SBP operators. Unlike previous methods based on CAS techniques, the new algorithm is guaranteed to generate *all* solutions in finite time. As an example, the algorithm is used to show the existence of standard diagonal-norm SBP operators with interior order as high as 18.

*Keywords:* high-order finite difference methods; summation by parts; diagonal energy norm

---

## 1. Introduction

The need for high-order numerical methods in the simulation of long-distance advection and wave propagation is well established. The introduction to a 1972 paper by Kreiss and Oliger [6], for example, provides a review of works dating back to the mid-1960's indicating that the inadequacy of second-order methods for applications in meteorology and oceanography was becoming apparent even then. The argument is based on the observation that the error in a numerical solution of a hyperbolic partial differential equation (PDE) scales roughly as a constant times  $T\omega^{(p+1)}h^p$ , where  $p$  is the order of accuracy of the numerical derivative operator,  $h$  is the spatial step size,  $\omega$  is a characteristic frequency of the solution, and  $T$  is the final time to which the equation is to be solved. Thus, in order to maintain a given level of error, the spatial step size should be scaled as  $h \sim \omega^{-(p+1)/p} T^{-1/p}$ . For problems wherein  $\omega$  and  $T$  are very large, it is essential that  $p$  be large as well, to avoid the need for overly small  $h$ .

For spatial derivative approximations based on finite differences, the requirement of high-order accuracy presents a particular challenge; in general, an arbitrary high-order finite difference approximation cannot be incorporated into a standard time-stepping scheme in a way that produces a stable solver. Because of this, much research has been conducted in the search for stable and accurate finite difference schemes.

### 1.1. Summation by Parts operators

Among the various types of high-order finite difference operators, the Summation by Parts (SBP) operators are unique in that their construction incorporates the construction of a natural discrete energy norm that can be used to prove stability for PDE solvers. Numerical schemes based on SBP operators have proven effective in simulating a wide variety of physical phenomena, including fluid flow [10, 15, 16], elastic wave propagation [2, 9, 12], and orbiting binary black holes [11].

The basic idea behind SBP operators (see, e.g., References [4, 7, 13]) is straightforward. One seeks to build, simultaneously, a finite difference operator and an associated vector norm that mimic, in a semi-discrete setting, some continuum “energy estimate” for the PDE. The one-dimensional advection equation on a bounded interval provides the canonical example.

---

\*Corresponding author

Consider the PDE

$$u_t + u_x = 0 \quad x \in (0, 1), \quad t > 0 \quad (1)$$

with suitable initial and boundary conditions. The energy

$$\mathcal{E}_u(t) = \|u(t, \cdot)\|_{L^2}^2 = \int_0^1 u(t, x)^2 dx,$$

has the property that, for  $u$  solving Equation (1),  $\mathcal{E}_u$  satisfies

$$\frac{d\mathcal{E}_u}{dt} = 2 \int_0^1 u u_t dx = -2 \int_0^1 u u_x dx = - \int_0^1 \frac{\partial}{\partial x} u^2 dx = u(t, 0)^2 - u(t, 1)^2. \quad (2)$$

Now, consider the following semi-discrete form of Equation (1). Let  $\{x_i\}_{i=1}^n$  be the grid of  $n$  equispaced nodes in  $[0, 1]$  with step size  $h = 1/(n-1)$ , and let  $v(t)$  be the  $n$ -vector approximating  $u$  in the method-of-lines interpretation. That is,  $v_i(t) \approx u(t, x_i)$  solves the semi-discrete equation

$$v_t + D_h v = 0 \quad (3)$$

for some  $n \times n$  finite difference operator  $D_h$ . Emulating the continuum case, let  $P_h$  be an  $n \times n$  symmetric positive definite matrix, and define the energy

$$E_v(t) = \|v(t)\|_P^2 = v(t)^T P_h v(t).$$

If  $P_h$  and  $D_h$  together satisfy the condition

$$P_h D_h + D_h^T P_h = e_n e_n^T - e_1 e_1^T = Q, \quad (4)$$

then it is straightforward to check that, with  $v$  a solution to Equation (3), the energy satisfies

$$\frac{dE_v}{dt} = v_t^T P_h v + v^T P_h v_t = -v^T (P_h D_h + D_h^T P_h) v = v_1^2 - v_n^2,$$

which is a semi-discrete analog of Equation (2). This property can be used to prove the stability of the fully discrete numerical solver.

Thus, the construction of an SBP first derivative operator (actually, an operator/norm pair) consists of constructing  $n \times n$  matrices  $D_h$  and  $P_h$  with the following properties.

- (P1)  $D_h$  is a finite difference approximation of the first derivative.
- (P2)  $P_h$  and  $D_h$  together satisfy the energy condition (4).
- (P3)  $P_h$  is a positive definite matrix.

### 1.2. SBP operators with diagonal norm

The remainder of the paper is restricted to the case that  $P_h$  is diagonal and that  $D_h$  coincides with a centered finite difference approximation sufficiently far from the boundary of the computational domain. The restriction to diagonal  $P_h$  is quite natural due to the fact that these are the only SBP operators for which standard techniques exist for proving stability for PDEs with variable coefficients or on multi-dimensional curvilinear grids [14]. Although SBP operators with non-diagonal (block) norm have recently been shown to be stabilizable on curvilinear grids [8], the question of existence of diagonal-norm SBP operators remains an interesting open problem; no such operators with interior order greater than 10 exist in the literature.

The requirement that  $D_h$  coincide with a centered difference on the interior is also natural, and is assumed throughout the literature. In this way, it is possible to assume that  $D_h$  and  $P_h$  have a simple dependence

on  $h$ . As an example, consider the following  $D_h$  and  $P_h$  with  $D_h$  equal to a 4th-order centered difference method in the interior.

$$D_h = \frac{1}{h} \begin{bmatrix} d_{11} & d_{12} & d_{13} & d_{14} & & & & & & \\ d_{21} & 0 & d_{23} & d_{24} & & & & & & \\ d_{31} & d_{32} & 0 & d_{34} & d_{35} & & & & & \\ d_{41} & d_{42} & d_{43} & 0 & d_{45} & d_{46} & & & & \\ \hdashline & & & \frac{1}{12} & -\frac{2}{3} & 0 & \frac{2}{3} & -\frac{1}{12} & & \\ & & & & \frac{1}{12} & -\frac{2}{3} & 0 & -\frac{2}{3} & -\frac{1}{12} & \\ & & & & & \frac{1}{12} & -\frac{2}{3} & 0 & \frac{2}{3} & -\frac{1}{12} \\ & & & & & & \ddots & \ddots & \ddots & \ddots \\ & & & & & & & \ddots & \ddots & \ddots \end{bmatrix} \quad (5)$$

$$P_h = h \operatorname{diag}(p_1, p_2, p_3, p_4, 1, 1, 1, \dots)$$

Note that only the top-left corners of  $D_h$  and  $P_h$  are included. The bottom-right corners can be reconstructed from the relationships

$$d_{n+1-i, n+1-j} = -d_{ij}, \quad p_{n+1-i} = p_i. \quad (6)$$

There are several important facts to observe about Equation (5). First, as specified,  $P_h$  is a diagonal matrix, and  $D_h$  coincides with a centered finite difference approximation for the rows  $i$  satisfying  $5 \leq i \leq n-4$ . This allows the corresponding diagonal entries of  $P_h$  to be taken as  $h$ . Observe also that the assumption that  $P_h$  is diagonal, combined with property (P2) gives the element-wise equations

$$p_i d_{ij} + p_j d_{ji} = \delta_{in} \delta_{jn} - \delta_{i1} \delta_{j1} = \begin{cases} -1 & \text{if } i = j = 1 \\ 1 & \text{if } i = j = n \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

Clearly, all diagonal entries of  $D_h$  other than the first and last must equal zero. Additionally, since  $d_{ij} \neq 0$  if and only if  $d_{ji} \neq 0$ , exactly which  $d_{ij}$  with  $i \leq 4$  and  $j \geq 5$  are nonzero can be determined by identifying the corresponding nonzero entries  $d_{ji}$ —for example, the entry  $d_{25}$  *must* equal zero because  $d_{52} = 0$ , while  $d_{45}$  *must not* be zero because  $d_{54} = -\frac{2}{3}$ . In this example, this leaves 20 unknown values to be determined—16 in  $D_h$  and 4 in  $P_h$ .

### 1.3. Equations and unknowns

The present paper considers the question of existence of a three-parameter class of SBP operators, of which Equation (5) is an example. The three parameters— $s$ ,  $o$  and  $r$ —describe both the shape and approximation accuracy of the first derivative operator. The parameter  $s$  defines the interior order of the operator: away from the boundary,  $D_h$  should behave as a  $2s$ -order centered difference operator. The parameter  $r$  defines the closure dimension:  $D_h$  corresponds with the centered difference for all rows  $i$  satisfying  $r+1 \leq i \leq n-r$ . From these two parameters alone, one can determine the “shape” of all possible derivative operators  $D_h$ . For example, Equation (5) gives the shape of the operator with  $s = 2$  and  $r = 4$ . The final parameter,  $o$ , gives the order of  $D_h$  near the boundary. That is,  $D_h$  acts as an  $o$ th order first derivative operator at the first  $r$  and last  $r$  grid points. In the literature, it is common to choose  $o = s$ , but we prefer to leave  $o$  free for more generality.

Equation (5) also provides some insight into counting the number of unknowns to be determined. Of the top  $r \times r$  block of  $D_h$ , there are  $r-1$  zero-valued diagonal entries, leaving  $r^2 - r + 1$  unknown values to be determined there. To count the unknowns in the remainder of the top  $r$  rows (i.e., nonzero  $d_{ij}$  with  $1 \leq i \leq r$ ,  $j > r$ ), one may equivalently count the number of nonzero entries in the first  $r$  columns below the  $r$ th row (i.e., nonzero  $d_{ij}$  with  $i > r$ ,  $1 \leq j \leq r$ ). For  $i = r+1, r+2, \dots, r+s$  there are respectively  $s, s-1, \dots, 1$  nonzero  $d_{ij}$  with  $j \leq r$ . This gives an additional  $s(s+1)/2$  unknowns. Finally, there are  $r$  unknown values in  $P_h$ , giving the full count

$$\# \text{unknowns} = r^2 + \frac{s(s+1)}{2} + 1.$$

These unknowns must be chosen in such a way that properties (P1)–(P3) are satisfied. First, in order to satisfy (P1), each of the first  $r$  rows of  $D_h$  must behave as an  $o$ th order finite difference operator. Since it is equivalent to differentiating all monomials up to degree  $o$  exactly, this condition amounts to solving a system of  $o + 1$  linear equations for each row  $d_i$  of  $D_h$  with  $i \leq r$ :

$$h d_i \begin{bmatrix} 1 & 1-i & (1-i)^2 & \dots & (1-i)^o \\ 1 & 2-i & (2-i)^2 & \dots & (2-i)^o \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & n-i & (n-i)^2 & \dots & (n-i)^o \end{bmatrix} = [0, 1, 0, \dots, 0]. \quad (8)$$

Next, condition (P2) can be written in component form as in Equation (7). Counting only the nontrivial equations (i.e., those for which  $d_{ij}$  and  $d_{ji}$  are not already known to equal zero) shows that there is one such equation for each pair of unknowns  $d_{ij}$  and  $d_{ji}$  with  $i \neq j$ , plus one more equation involving  $d_{11}$ . Thus,

$$\# \text{equations} = r(o+1) + \frac{r(r-1)}{2} + \frac{s(s+1)}{2} + 1.$$

Finally, property (P3) is equivalent to the requirement that  $p_i > 0$  for  $i = 1, 2, \dots, r$ , giving  $r$  inequality constraints. This paper considers the questions of existence and construction of an SBP operator/norm pair satisfying properties (P1)–(P3) for a particular choice of  $s$ ,  $o$  and  $r$ . That is, we suppose that  $s$ ,  $o$  and  $r$  are some fixed values and attempt to determine whether the equations and inequalities described above can be simultaneously satisfied. We shall often use the phrases “the SBP operator exists” or “no SBP operator exists”. This should be understood as addressing the question of existence of an SBP operator/norm pair satisfying the above conditions for this particular choice of parameters.

#### 1.4. Solving by Computer Algebra System

In the literature, it appears that—although not all papers describe exactly the techniques employed—the standard method for solving the equations described in Section 1.3 is to input the equations into a Computer Algebra System (CAS) and attempt to solve them by automatic symbolic manipulation. The difficulty with this technique lies in the inability to state with certainty that all solutions are found. Indeed, Equation (7) is a nonlinear system of equations, and there are no known algorithms for symbolically finding all solutions to a general nonlinear system. Hence, it is desirable to have a more robust solution method that, given sufficient memory and time, will produce *all* solutions or report correctly that none exists. This is the purpose of the present work.

#### 1.5. Overview

The paper is arranged as follows. In Section 2, the nonlinear SBP equations are transformed into a sequence of *linear* problems, each of which can be solved in finite time with certainty, thus providing an algorithm to decide the existence of an SBP operator with given parameters. Section 3 describes several novel results obtained by the new algorithm, and provides an example of a method for producing “useful” SBP operators. Finally, Section 4 summarizes our conclusions and states several open problems. We have also included, in Appendix A, a demonstration of the behavior of the new algorithm, step by step, for a simple example. The coefficients for two new SBP operators (of 6th and 7th order respectively) are included online as supplementary data, as described in Appendix B.

## 2. Algorithm for existence

This section describes a method for algorithmically deciding the SBP problem for given parameters  $s$ ,  $o$  and  $r$ . That is, the algorithm presented here determines whether or not such an operator/norm pair exists, but does not completely construct one. The construction, assuming existence is known, is postponed until the next section. The reader may find it helpful to refer periodically to Appendix A for a concrete example of each step of the following procedure.

In terms of implementation, the entire process that follows should be performed in exact arithmetic (i.e., using rational numbers), as the result relies upon the *exact* solution of linear problems. For the results presented in this paper, the Python library `sympy` [17] was used for exact arithmetic. It remains an open problem whether a similar algorithm is possible in finite precision (see Section 4).

### 2.1. Setup

To initialize the algorithm, the finite difference weights for the interior are computed exactly. Next, an initial sweep of the upper-left  $r \times (r + s)$  block of  $D_h$  is performed, identifying all unknowns in the first  $r$  rows, as described in Section 1.3.

### 2.2. Reduce unknowns

The next step of the process is to reduce the number of unknowns by enforcing property (P1). Since this amounts to solving a linear system of equations (Equation (8)) for each of the first  $r$  rows of  $D_h$ , every solution can be found. More specifically, let  $d_i$  be a vector of the unknowns in row  $i$  of  $D_h$ —that is,  $d_i$  is made up of any of the  $d_{ij}$  not already known to be zero. Assuming sufficiently many degrees of freedom exist in  $d_i$  for an  $\ell_i$  order derivative approximation,  $d_i$  can be written in the form

$$d_i = d_i^0 + \sum_{k=1}^{\ell_i} w_{ik} d_i^k,$$

where  $d_i^0$  is a derivative approximation and the  $w_{ik}$  parametrize the  $\ell_i \geq 0$  additional degrees of freedom in the directions  $d_i^k$ , if any exist. In the case  $\ell_i = 0$ , the equation should be interpreted as  $d_i = d_i^0$ . Written in component form, this set of equations becomes

$$d_{ij} = d_{ij}^0 + \sum_{k=1}^{\ell_i} w_{ik} d_{ij}^k \quad (9)$$

where  $i \leq j$  range over all unknown entries in the  $D_h$  matrix. In our code, we have used the (symbolic) function `rref` to row-reduce the linear systems and identify both a particular solution  $d_i^0$  as well as the vectors  $d_i^k$  spanning the nullspace.

### 2.3. Construct the $w$ -system

Now consider property (P2), which requires that Equation (7) hold for all  $i \leq j$  ranging over the unknowns in  $D_h$ . Substituting Equation (9) into Equation (7) results in the equations

$$p_i \left( d_{ij}^0 + \sum_{k=1}^{\ell_i} w_{ik} d_{ij}^k \right) + p_j \left( d_{ji}^0 + \sum_{k=1}^{\ell_j} w_{jk} d_{ji}^k \right) = \delta_{in} \delta_{jn} - \delta_{i1} \delta_{j1}. \quad (10)$$

It is important to observe that the rows beyond  $i = r$  are fully determined; they have no degrees of freedom. Thus, when  $j > r$  Equation (10) only contains unknowns associated with  $d_{ij}$  since  $d_{ji}$  is known. Defining the vector  $w$  from an enumeration of the unknowns  $w_{ik}$ , and  $p = (p_1, p_2, \dots, p_r)^T$ , the above equations can be written in the form

$$A_p w = b_p, \quad (11)$$

where  $A_p$  is a matrix depending only on  $p$  and  $b_p$  is a vector also depending only on  $p$ . This proves the following lemma

**Lemma 1.** *The SBP operator exists if and only if Equation (11) is solvable for  $p$  and  $w$  with  $p$  strictly positive.*

#### 2.4. Construct the $p$ -system

The reduction provided by Lemma 1 is interesting in that it reduces the number of degrees of freedom to only those present in  $p$  and  $w$ . However, much more can be said. First, consider the application of the standard solvability alternative of linear algebra to Equation (11). In this setting, it can be stated as the following lemma.

**Lemma 2.** *Let  $Z_p$  be a matrix (parametrized by  $p$ ) whose columns form a basis for the nullspace  $N(A_p^T)$ . The SBP operator exists if and only if*

$$Z_p^T b_p = 0$$

*has a solution  $p$  with strictly positive entries.*

The lemma essentially reduces the existence question to the question of solvability of a (generally non-linear) system of equations of  $p$  alone; the unknown vector  $w$  has been excluded. Once again, this is an interesting result, but, by now invoking the special diagonal structure of  $P_h$ , we arrive at an even stronger result.

Observe from Equation (10) and Equation (11) that the entries of  $b_p$  are affine functions of the  $p_i$ . Moreover, since each  $w_{ik}$  only appears in terms multiplied by  $p_i$ , each column of  $A_p$  can be written as a real vector scaled by one of the  $p_i$ —that is,  $A_p = A\Lambda_p$ , where  $A$  is a fixed matrix independent of  $p$ , and  $\Lambda_p$  is a diagonal matrix whose entries are elements of the  $p_i$  (see Equation (A.2), for an example). Thus, Equation (11) can be written in the form

$$A\Lambda_p w = b_0 + Bp. \quad (12)$$

If  $P_h$  is positive definite, then  $\Lambda_p$  is invertible, so the range of  $A_p$  is equal to the range of  $A$ . This implies the following theorem.

**Theorem 1.** *Let  $Z$  be a matrix whose columns form a basis for  $N(A^T)$ . The SBP operator exists if and only if the system*

$$Z^T Bp = -Z^T b_0 \quad (13)$$

*has a solution  $p$  with strictly positive entries.*

This is the key result of the present contribution, as it ties the solvability of Equation (13) with positive  $p$  directly to the existence of the SBP operator.

#### 2.5. Solve the LP problem

The preceding steps have all involved very simple linear algebra manipulations (primarily substitution and Gaussian elimination), and have reduced the SBP existence problem to the problem of solving a linear system with simple inequality constraints. The most complicated step in seeking the solution to this final equation (13) is exactly in enforcing these constraints. To see how this problem can be solved algorithmically, consider the following optimization problem

$$\begin{aligned} & \text{maximize} && \min_i p_i \\ & \text{subject to} && Z^T Bp = -Z^T b_0 \end{aligned} \quad (14)$$

and note that Equation (13) has a strictly positive solution if and only if the value of the optimization problem is strictly positive (with the understanding that the value of Equation (14) is  $-\infty$  if there are no feasible points  $p$ ). The optimization problem can be algorithmically solved through the following manipulations.

First, it is convenient to determine whether Equation (13) is even solvable—which can again be accomplished by `rref`. If it is not, then we may state with certainty that no SBP operator exists. Now, assume that Equation (13) is solvable. If there is a unique solution,  $p$ , then the existence of the SBP operator can be immediately determined by checking whether this  $p$  has strictly positive entries. Finally, consider the case that a solution exists, but is non-unique. In this case, the solution set may be parametrized as

$$p = p_0 + Gq$$

for some particular solution  $p_0$  and some  $G$  whose columns span the nullspace of  $Z^T B$ . Equation (14) can now be treated by a common trick that transforms the problem into a standard linear program (LP):

$$\begin{aligned} & \text{maximize} && t \\ & \text{subject to} && p_i \geq t, \quad i = 1, 2, \dots, r \\ & && p = p_0 + Gq \end{aligned} \tag{15}$$

As an LP, Equation (15) can be solved by the simplex method, thus providing an algorithm for solving Equation (13). In this case, we conclude that the SBP operator exists if and only if the solution to Equation (15) is positive. From an implementation perspective, this is the most complex step, as it requires a simplex solver in *exact arithmetic*. We did not find an existing library for this, and so implemented our own simplex solver in Python.

### 3. Existence and nonexistence of SBP operators

This section presents some novel results based on the algorithm of the previous section. It is worth remarking that, although the algorithm is provably correct, the following computational results rely on “computer-assisted proof”. The numerators and denominators of the rational numbers involved are sufficiently large that we cannot hope to perform the Gaussian elimination and simplex method steps by hand except in a small number of cases. For example, the value for  $p_1$  in the case  $s = o = 8$ ,  $r = 23$  is

$$p_1 = \frac{83852077150009258297147}{299027329581685985280000}.$$

Although we have done our best to test our code and to compare with SBP results in the literature, the results presented in this text are nevertheless vulnerable to errors either in the `sympy` rational number manipulation routines, the Gaussian elimination routine or in the simplex solver. As an example, an earlier version of the code produced incorrect results from time to time due to some unexpected behavior in the symbolic operations of a particular commercial software tool. We have a high degree of confidence in the computational results presented in this paper, but certainly encourage their verification by others.

#### 3.1. The smallest $r$ for $o = s$

We first consider the case of an SBP operator of order  $2s$  in the interior and  $s$  on the boundary. It can be readily seen that if a solution exists for a particular choice of  $s$ ,  $o$ , and  $r$ , then this is also a solution for  $s$ ,  $o$  and any larger  $r$ . Thus, it is possible to perform a bisection search to locate the smallest  $r$  for which an SBP operator with the given choice  $s = o$  exists. Table 1 presents the results of this parameter sweep. The table also gives the dimension of the solution space of  $P_h$ , the value of Equation (14), and the dimension of the solution space of  $D_h$  for the optimal  $P_h$ .

To our knowledge, the results for  $s > 5$  are unknown in the literature; it was formerly unknown whether diagonal-norm SBP operators of these orders even existed. Also of interest is the nontrivial dependence of  $r$  on  $s$ . Based on previous results for the cases  $s = 2, 3$  and  $4$ , it might be expected that SBP operators exist for all  $s = o = r$ . However, by the nature of the parameter sweep conducted here, we conclude that, for a given  $s$ , no SBP operator (with  $o = s$ ) exists for  $r$  smaller than the value given in the table. The case  $s = 5$ ,  $r = 11$  has already been reported (see, e.g., References [5, 8]). However, while a footnote of Reference [5] states that the choice  $s = 5$ ,  $r = 10$  “did not result in a positive definite norm”, no proof or explanation is given.

#### 3.2. Optimization of the derivative operator

Another interesting observation about the results presented in Table 1 is that the number of degrees of freedom in  $D_h$  grows rapidly with increasing  $s$ . At the end of the algorithm of the previous section we are generally left, not with a single SBP operator, but with an entire linear manifold:

$$D_h = D_0 + \sum_j \xi_j D_j \tag{16}$$

$s$	$r$	dof $P_h$	dof $D_h$	$\min_i p_i$
1	1	0	0	5.000e-01
2	4	0	0	3.541e-01
3	6	0	1	3.159e-01
4	8	0	3	2.575e-01
5	11	1	10	2.077e-01
6	14	2	21	9.683e-03
7	19	5	55	1.907e-01
8	23	7	91	4.652e-02
9	28	10	—	4.622e-02

Table 1: For the case  $o = s$ , the table reports the *smallest* value of  $r$  for which an SBP operator exists. When  $P_h$  is non-unique, the number of degrees of freedom in  $P_h$  is reported as “dof  $P_h$ ”. The value of Equation (14) is reported as  $\min_i p_i$ . For the optimal  $P_h$ , the degrees of freedom of  $D_h$  is reported as “dof  $D_h$ ”. Due to insufficient memory, it was not possible to count the number of degrees of freedom in  $D_h$  for the  $s = 9$  case.

where  $j$  varies through all degrees of freedom. Although the primary concern of the present paper is the *existence* of SBP operators, the question of “Which is best?” is also important. As described in Reference [5], there are a wide range of options for defining “best”. Rather than consider each of these, we focus on a particular objective function: minimizing the spectral radius  $\rho(D_h)$ . This objective function is interesting because it controls the CFL condition for explicit PDEs solvers. Moreover, it is unique among the objective functions considered in the reference in that it is non-convex in the parameters  $\xi_j$  of  $D_h$ , and therefore difficult to optimize globally. Hence the remark in Reference [5]: “Therefore, when we refer to minimizing the spectral radius, we perform a numerical minimization and do not claim that we have actually found a global minimum.”

In this paper, we suggest an alternative to minimizing the spectral radius directly. The key point is that, although  $\rho(C)$  is not a convex function in the entries of  $C$  in general, it *is* convex for normal matrices  $C$ . And although  $D_h$  is not a normal matrix in general, it is close in some sense to a normal matrix, because

$$\left(P_h D_h - \frac{1}{2}Q\right) + \left(P_h D_h - \frac{1}{2}Q\right)^T = 0$$

(see Equation (4)). Since the surrogate matrix  $P_h D_h - \frac{1}{2}Q$  is skew-symmetric and therefore normal, its spectral radius agrees with its operator 2-norm and, thus, is a convex function of its entries. Moreover, in this norm

$$\rho(D_h) \leq \|D_h\| \leq \|P_h^{-1}\| \cdot \|P_h D_h\| \leq \|P_h^{-1}\| \cdot \left(\|P_h D_h - \frac{1}{2}Q\| + \frac{1}{2}\right).$$

Provided  $\|P_h\|$  is not too large, minimizing the norm of the surrogate matrix tends to make the spectral radius of  $D_h$  small. Defining

$$C_0 = P_h D_0 - \frac{1}{2}Q, \quad C_i = P_h D_i, \quad C(\xi) = C_0 + \sum_j \xi_j C_j$$

the goal is to minimize  $\|C(\xi)\|$  with respect to  $\xi = (\xi_j)$ . It turns out that this problem can be easily transformed into a Semidefinite Program (SDP) [3, Sec. 4.6.3], treatable by a number of standard solvers.

Our implementation of this idea is to choose a particular  $N$  (we chose  $N = 100$  for our examples) and to numerically minimize the norm of the surrogate  $\|C(\xi)\|$  with respect to  $\xi$ . Unlike in the previous section, there is no apparent need to perform this optimization in exact arithmetic. Instead, the elements of the  $C_i$  are evaluated in double precision and are used to set up the SDP, which is then solved through the `cvxopt` package [1]. Once the optimal  $\xi$  is found, the corresponding  $D_h$  is formed from Equation (16).

Using this technique on the case  $s = o = 7$ ,  $r = 19$  (with a 55-dimensional search space) we located an SBP operator such that  $\rho(D_h) \approx 2/h$ , as verified with several choices of  $h$ . Applying the same technique



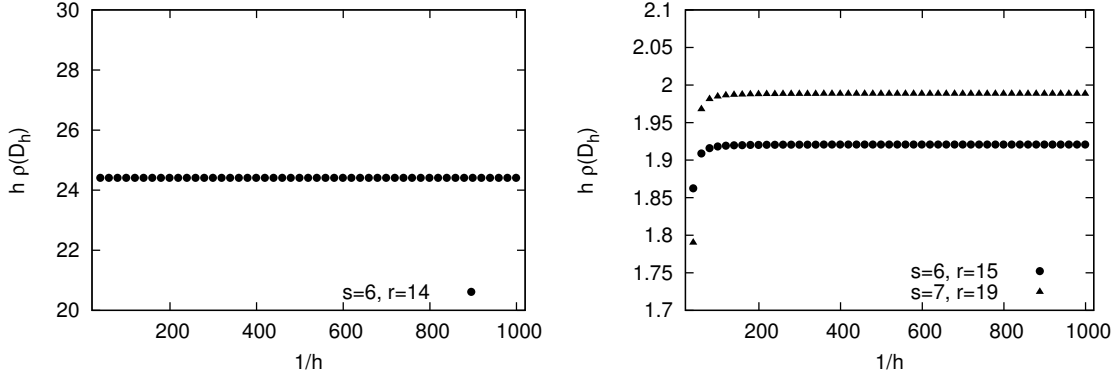


Figure 1: On the left, the spectral radius as a function of  $1/h$  for the optimized  $s = o = 6, r = 14$  SBP operator. On the right, the spectral radii as functions of  $1/h$  for the optimized  $s = o = 6, r = 15$  and  $s = o = 7, r = 19$  SBP operators.

to the case  $s = o = 6, r = 14$ , however, did not produce suitable results, as might be expected from careful inspection of Table 1. In particular, the value  $\min_i p_i$  associated with this case is a very small number, implying that  $\|P_h^{-1}\|$  is large. When we attempted to minimize the surrogate matrix in this case the resulting SBP operator exhibited (numerically) the scaling  $\rho(D_h) \approx 24/h$ —significantly larger than one might wish. This problem can be remedied as follows.

Recall that if the SBP equations are solvable for  $s, o$  and  $r$ , then they are solvable for the same  $s$  and  $o$  with any larger  $r$ . In general, increasing  $r$  leads to a larger number of degrees of freedom in both  $P_h$  and  $D_h$ . So, it is reasonable to ask whether choosing  $r > 14$  might improve the result in the case  $s = 6$ . With the choice  $s = 6, r = 15$  there is an SBP operator with  $\min_i p_i \approx 0.24$ . In this case, the resulting SBP operator exhibits the scaling  $\rho(D_h) \approx 1.92/h$ . Thus, we conclude, that the smallest  $r$  for which an SBP operator exists is not necessarily the best choice of  $r$ . Apparently, it is useful to choose an  $r$  for which  $\|P_h^{-1}\|$  is not too large. It is interesting to note that, although the objective function in Equation (14) was not chosen specifically for this property, a side-effect of the algorithm described in this paper is to choose, among all possible  $P_h$ , the one with smallest  $\|P_h^{-1}\|$ .

Figure 1 shows the numerically computed spectral radii of the operators described in this section as functions of  $1/h$ . The coefficients for the optimized  $s = 6$  ( $r = 15$ ) and  $s = 7$  operators are included online as supplementary data for this paper, as described in Appendix B.

### 3.3. Operators with $s \geq o$

As a final application of the algorithm, we consider the case that  $s \neq o$ —that is, the closure has different order than is usually specified for diagonal-norm SBP operators. Although we do not have a proof of this fact, we have not yet obtained existence of an SBP operator with  $o > s$ , and it is possible that no such operator exists. On the other hand, it is easy to find SBP operators with  $o < s$ . Table 2 gives a summary of some results in this regard. For this test, we allowed  $o$  to range from  $s - 2$  to  $s$  and found the smallest  $r$  for which the SBP operator with these parameters exist. For this  $r$  and the two subsequent values of  $r$ , the table summarizes the number of degrees of freedom in  $P_h$  as well as the value  $\min_i p_i$ . It is interesting to note that, if an SBP operator exists for a particular value of  $s, o$  and  $r$ , then an SBP operator also appears to exist with the same  $o$  and  $r$  with any larger value of  $s$ , and that  $\min_i p_i$  seems to primarily depend on  $o$  and  $r$  alone (although a comparison of the cases  $(s, o, r) = (4, 4, 8)$  and  $(5, 4, 8)$  shows that there is at least some  $s$  dependence). Finally, it appears that the number degrees of freedom in  $P_h$  is equal to  $r - 2o$ .

## 4. Conclusion and future research

This paper introduces an algorithm that provably answers the question of existence of diagonal-norm SBP operators parametrized by  $s, o$  and  $r$ . We consider this an improvement on the traditional method of

$s$	$o$	$r$	dof $P_h$	$\min_i p_i$	$s$	$o$	$r$	dof $P_h$	$\min_i p_i$
4	2	4	0	3.486e-01	6	4	8	0	2.570e-01
4	2	5	1	4.842e-01	6	4	9	1	3.121e-01
4	2	6	2	5.606e-01	6	4	10	2	3.367e-01
4	3	6	0	3.156e-01	6	5	11	1	2.076e-01
4	3	7	1	3.598e-01	6	5	12	2	2.997e-01
4	3	8	2	4.413e-01	6	5	13	3	3.167e-01
4	4	8	0	2.575e-01	6	6	14	2	9.683e-03
4	4	9	1	3.121e-01	6	6	15	3	2.403e-01
4	4	10	2	3.367e-01	6	6	16	4	2.992e-01
5	3	6	0	3.156e-01	7	5	11	1	2.076e-01
5	3	7	1	3.598e-01	7	5	12	2	2.997e-01
5	3	8	2	4.413e-01	7	5	13	3	3.167e-01
5	4	8	0	2.570e-01	7	6	14	2	9.682e-03
5	4	9	1	3.121e-01	7	6	15	3	2.403e-01
5	4	10	2	3.367e-01	7	6	16	4	2.992e-01
5	5	11	1	2.077e-01	7	7	19	5	1.907e-01
5	5	12	2	2.997e-01	7	7	20	6	2.923e-01
5	5	13	3	3.167e-01	7	7	21	7	3.012e-01

Table 2: For several cases with  $o \leq s$ , the table reports the *smallest* value of  $r$  for which an SBP operator exists as well as the subsequent two  $r$  values. When  $P_h$  is non-unique, the number of degrees of freedom in  $P_h$  is reported as “dof  $P_h$ ”. The value of Equation (14) is reported as  $\min_i p_i$ .

using a CAS to solve the nonlinear system of equations and inequalities, which may not locate all solutions; the new algorithm can provably answer with certainty in finite time whether or not an SBP operator with given parameters exists. To our knowledge, this is the first such algorithm for SBP operators. Our hope is that this methodological approach will lead to further research in the field, and to this end, we conclude with a list of what we consider to be interesting directions of future research.

*Floating point algorithms.* As remarked in Section 2, the current algorithm relies crucially on the use of exact arithmetic. Although this is a significant improvement over the use of symbolic solvers, it is still limited by the need to store and manipulate rational numbers. This appears to be the principal bottleneck preventing the discovery of even higher-order SBP operators than those presented in this text, since this representation leads to very large memory requirements and computationally expensive arithmetic operations. It would be interesting to know if there exists a similar algorithm for finding (approximate) SBP operators in floating point arithmetic. Even the use of a variable-precision library would be an improvement over the need for rational numbers.

*Alternative energy norms.* The algorithm described in Section 2.5 chooses, among all possible  $P_h$ , the one that maximizes  $\min_i p_i$ . This choice is useful for two reasons. First, if the optimal value is non-positive, then we immediately conclude that no SBP operator exists. Moreover, as remarked in Section 3.2, this choice is good for the application of optimizing the spectral radius of the SBP derivative operator. On the other hand, if Equation (13) has a manifold of solutions and if one of these solutions is strictly positive, then, in fact, Equation (13) has an infinite number of strictly positive solutions. It is not clear how the choice of  $P_h$  might influence later steps of the algorithm.

*Second derivative approximations.* An important component of SBP-based solvers is the implementation of a second-derivative operator that interacts correctly with the first derivative operator and norm. The question of incorporating second derivative approximations into the new algorithm is a subject of ongoing investigation.

*Compact stencil sizes.* As can be seen from Table 1, the closure size,  $r$ , appears to grow rapidly as  $s$  increases. For example, the 18th order ( $s = 9$ ) centered operator requires at least 28 points for a 9th order closure. This derivative operator contains a 9th order derivative approximation with a stencil width of  $28 + 9 = 37$ . It is not clear whether a method with such a wide stencil would actually be useful in applications, and it would be interesting to know if there are generalizations of the SBP framework that can reduce this size.

*Algorithms for block norms.* The assumption of diagonal norm  $P_h$  played an important role in constructing the *linear* system of Theorem 1, which allows all solutions to be found algorithmically. In principle, the same type of reduction of unknowns as in Section 2.2 can be applied to the general (block norm) SBP construction, leading to a system of equations of the form of Equation (11). However, in the block norm case,  $A_p$  cannot be decomposed as nicely as in Equation (12). Nevertheless, Lemma 2 still holds, effectively allowing the reduction of the SBP problem to a *nonlinear* system in  $p$  alone. This observation could prove useful in constructing such operators, and perhaps an algorithm could be devised for this particular form of nonlinear system.

## Acknowledgments

Joshua Klarmann's work on this project was sponsored by the McNair Scholars' Program and supported by the National Science Foundation under Award No. EPS-0903806 and matching support from the State of Kansas through the Kansas Board of Regents; further funding was received from a scholarship provided by the College of Arts and Sciences at Kansas State University.

## Appendix A. The algorithm in action

As an example of the algorithm presented in Section 2, consider the question of existence of an SBP operator with  $s = o = 1$  and  $r = 3$ . Assuming  $P_h = h \operatorname{diag}(p_1, p_2, p_3, 1, 1, \dots)$ , the derivative matrix  $D_h$  must have the form

$$D_h = \frac{1}{h} \begin{bmatrix} d_{11} & d_{12} & d_{13} & & & & \\ d_{21} & 0 & d_{23} & & & & \\ d_{31} & d_{32} & 0 & d_{34} & & & \\ & & -\frac{1}{2} & 0 & \frac{1}{2} & & \\ & & & -\frac{1}{2} & 0 & \frac{1}{2} & \\ & & & & \ddots & \ddots & \ddots \end{bmatrix} \quad (\text{A.1})$$

*Reduce unknowns.* After satisfying (P1) on the first 3 rows, the remaining number of degrees of freedom for this example are  $\ell_i = 1, 0$  and  $1$ , for  $i = 1, 2$ , and  $3$  respectively. Writing  $d_1$  as the row vector  $(d_{11}, d_{12}, d_{13})$ , Equation (9) in this case is

$$d_1 = (-1, 1, 0) + w_{11} (1, -2, 1);$$

all first-order finite difference approximations of  $f'(x_0)$  on the nodes  $x_0, x_1, x_2$  take the form

$$h f'(x_0) \approx -(1 - w_{11}) f(x_0) + (1 - 2w_{11}) f(x_1) + w_{11} f(x_2).$$

Similarly, writing  $d_2 = (d_{21}, d_{23})$ , Equation (9) is

$$d_2 = \left( -\frac{1}{2}, \frac{1}{2} \right),$$

which encodes the representation of all first-order finite difference approximations (actually second-order in this case) of  $f'(x_0)$  on the nodes  $x_{-1}, x_1$  as

$$h f'(x_0) \approx \frac{f(x_1) - f(x_{-1}))}{2}.$$

Finally, writing  $d_3$  as the row vector  $(d_{31}, d_{32}, d_{34})$ , Equation (9) becomes

$$d_3 = (-1, 1, 0) + w_{31} (2, -3, 1).$$

Note that all first-order finite difference approximations of  $f'(x_0)$  on the nodes  $x_{-2}, x_{-1}, x_1$  take the form

$$h f'(x_0) \approx -(1 - 2w_{31}) f(x_{-2}) + (1 - 3w_{31}) f(x_{-1}) + w_{31} f(x_1).$$

*Construct the  $w$ -system.* Continuing the example of Equation (A.1), Equation (10) shows that, for example,

$$p_1 w_{11} = -\frac{1}{2} + p_1, \quad p_1 w_{11} + 2p_3 w_{31} = p_3, \quad p_3 w_{31} = \frac{1}{2}, \quad \text{etc.}$$

Written as in Equation (12), the full system for the  $p_i$  and  $w_{ik}$  is

$$\begin{bmatrix} 1 & 0 \\ -2 & 0 \\ 1 & 2 \\ 0 & -3 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} p_1 & 0 \\ 0 & p_3 \end{bmatrix} \begin{bmatrix} w_{11} \\ w_{31} \end{bmatrix} = \begin{bmatrix} -\frac{1}{2} \\ 0 \\ 0 \\ 0 \\ \frac{1}{2} \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 \\ -1 & \frac{1}{2} & 0 \\ 0 & 0 & 1 \\ 0 & -\frac{1}{2} & -1 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix} \quad (\text{A.2})$$

*Construct the  $p$ -system.* In this case, the matrix  $Z$  whose columns span the nullspace of  $A^T$  is

$$Z = \begin{bmatrix} 2 & -\frac{3}{2} & \frac{1}{2} \\ 1 & 0 & 0 \\ 0 & \frac{3}{2} & -\frac{1}{2} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Thus, the system in Equation (13) is

$$\begin{bmatrix} 1 & \frac{1}{2} & 0 \\ -\frac{3}{2} & -\frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & -\frac{1}{2} \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix} = \begin{bmatrix} 1 \\ -\frac{3}{4} \\ -\frac{1}{4} \end{bmatrix}.$$

The SBP operator exists if and only if a strictly positive solution vector exists.

*Solve the LP problem.* The linear system in  $p$  has a 1-dimensional solution manifold

$$p(q) = \begin{bmatrix} -\frac{1}{2} + q \\ 3 - 2q \\ q \end{bmatrix}.$$

This  $p(q)$  is positive if and only if  $1 < 2q < 3$ , so certainly the SBP operator exists. Moreover, the minimum value of  $p(q)$  is maximized when  $q = 7/6$  giving  $p = (\frac{2}{3}, \frac{2}{3}, \frac{7}{6})^T$ .

With this choice of norm  $P_h$ , there is a unique SBP derivative operator  $D_h$ :

$$D_h = \frac{1}{h} \begin{bmatrix} -\frac{3}{4} & \frac{1}{2} & \frac{1}{4} & \vdots & & & \\ -\frac{1}{2} & 0 & \frac{1}{2} & \vdots & & & \\ -\frac{1}{2} & -\frac{2}{7} & 0 & \frac{3}{7} & & & \\ \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \ddots \end{bmatrix}.$$

## Appendix B. Coefficients of new SBP operators

The coefficients for the new 6th- and 7th-order SBP operators described in Section 3.2 are included online as text files. The files `P_6_6_15.txt` and `D_6_6_15.txt` hold the coefficients for  $P_h$  and  $D_h$  of the 6th order method, respectively. While `P_7_7_19.txt` and `D_7_7_19.txt` hold the coefficients for  $P_h$  and  $D_h$  of the 7th order method.

The coefficients are scaled to the case  $h = 1$ . In the case of the norm matrices  $P_h$ , the data are stored in rows of 2 columns. Each row holds a pair  $(i, v)$  indicating that  $p_i = v$ . Only the upper-left corner is given, since the lower-right corner can be obtained from Equation (6). For the files containing  $D_h$  coefficients, each row contains a triple  $(i, j, v)$  indicating that  $d_{ij} = v$ . Again, only the upper-left corner is given.

- [1] ANDERSEN, M. S., DAHL, J., AND VANDENBERGHE, L. *CVXOPT: A Python package for convex optimization, Version 1.1.6*, 2013. Available at <http://cvxopt.org>.
- [2] APPELÖ, D., AND PETERSSON, N. A. A stable finite difference method for the elastic wave equation on complex geometries with free surfaces. *Commun. Comput. Phys.* 5, 1 (2009), 84–107.
- [3] BOYD, S. P., AND VANDENBERGHE, L. *Convex optimization*. Cambridge university press, 2004.
- [4] CARPENTER, M. H., GOTTLIEB, D., AND ABARBANEL, S. Time-stable boundary conditions for finite-difference schemes solving hyperbolic systems: methodology and application to high-order compact schemes. *J. Comput. Phys.* 111, 2 (1994), 220–236.
- [5] DIENER, P., DORBAND, E. N., SCHNETTER, E., AND TIGLIO, M. Optimized high-order derivative and dissipation operators satisfying summation by parts, and applications in three-dimensional multi-block evolutions. *J. Sci. Comput.* 32, 1 (2007), 109–145.
- [6] KREISS, H.-O., AND OLIGER, J. Comparison of accurate methods for the integration of hyperbolic equations. *Tellus* 24, 3 (1972), 199–215.
- [7] KREISS, H.-O., AND SCHERER, G. Finite element and finite difference methods for hyperbolic partial differential equations. In *Mathematical Aspects of Finite Elements in Partial Differential Equations* (1974), Academic Press, pp. 195–212.
- [8] MATTSSON, K., AND ALMQUIST, M. A solution to the stability issues with block norm summation by parts operators. *J. Comput. Phys.* 253 (2013), 418–442.
- [9] NILSSON, S., PETERSSON, N. A., SJÖGREEN, B., AND KREISS, H.-O. Stable difference approximations for the elastic wave equation in second order formulation. *SIAM J. Numer. Anal.* 45, 5 (2007), 1902–1936.
- [10] OSUSKY, M., HICKEN, J. E., AND ZINGG, D. W. A parallel Newton-Krylov-Schur flow solver for the Navier-Stokes equations using the SBP-SAT approach. In *48th AIAA Aerospace Sciences Meeting and Exhibit, Orlando, Florida, AIAA-2010-116* (2010).
- [11] PAZOS, E., TIGLIO, M., DUEZ, M. D., KIDDER, L. E., AND TEUKOLSKY, S. A. Orbiting binary black hole evolutions with a multipatch high order finite-difference approach. *Phys. Rev. D* 80, 2 (2009), 024027.
- [12] SJÖGREEN, B., AND PETERSSON, N. A. A fourth order accurate finite difference scheme for the elastic wave equation in second order formulation. *J. Sci. Comput.* 52, 1 (2012), 17–48.
- [13] STRAND, B. Summation by parts for finite difference approximations for  $d/dx$ . *J. Comput. Phys.* 110, 1 (1994), 47–67.
- [14] SVÄRD, M. On coordinate transformations for summation-by-parts operators. *J. Sci. Comput.* 20, 1 (2004), 29–42.
- [15] SVÄRD, M., CARPENTER, M. H., AND NORDSTRÖM, J. A stable high-order finite difference scheme for the compressible Navier-Stokes equations, far-field boundary conditions. *J. Comput. Phys.* 225 (July 2007), 1020–1038.
- [16] SVÄRD, M., AND NORDSTRÖM, J. A stable high-order finite difference scheme for the compressible Navier-Stokes equations: No-slip wall boundary conditions. *J. Comput. Phys.* 227, 10 (2008), 4805 – 4824.
- [17] SYMPY DEVELOPMENT TEAM. *SymPy: Python library for symbolic mathematics*, 2014. Available at <http://www.sympy.org>.